Towards broadening Forensic DNA Phenotyping beyond pigmentation: Improving the prediction of head hair shape from DNA

Ewelina Pośpiech[1,2], Yan Chen[3,4], Magdalena Kukla-Bartoszek[5], Krystal Breslin[6], Anastasia Aliferi[7], Jeppe D. Andersen[8], David Ballard[7], Lakshmi Chaitanya[9], Ana Freire-Aradas[10,11], Kristiaan J. van der Gaag[12], Lorena Girón-Santamaría[11], Theresa E. Gross[10], Mario Gysi[13], Gabriela Huber[14], Ana Mosquera-Miguel[11], Charanya Muralidharan[6], Małgorzata Skowron[15], Ángel Carracedo[11,16], Cordula Haas[13], Niels Morling[8], Walther Parson[14,17], Christopher Phillips[11], Peter M. Schneider[10], Titia Sijen[12], Denise Syndercombe-Court[7], Marielle Vennemann[18], Sijie Wu[4,19], Shuhua Xu[4,19,20,21], Li Jin[19,20], Sijia Wang[4,19,20], Ghu Zhu[22], Nick G. Martin[22], Sarah E. Medland[22], EUROFORGEN_NoE Consortium, Wojciech Branicki[2], Susan Walsh[6], Fan Liu[3,4,9] and Manfred Kayser[9*]

[1]Institute of Zoology and Biomedical Research, Faculty of Biology and Earth Sciences, Jagiellonian University, Gronostajowa st. 9, 30-387, Kraków, Poland

[2]Malopolska Centre of Biotechnology, Jagiellonian University, Gronostajowa st. 7A, 30-387 Kraków, Poland

[3]Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beichen West Road 1-104, Chaoyang, Beijing, 100101, P.R. China

[4]University of Chinese Academy of Sciences, 19 Yuquan Road, Shijingshan, Beijing, 100049, P.R. China

[5]Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, Gronostajowa st. 7, 30-387, Kraków, Poland

[6]Department of Biology, Indiana University Purdue University Indianapolis (IUPUI), Indiana, USA

[7]King's Forensics, Faculty of Life Sciences and Medicine, King's College London, 150 Stamford

This is the author's manuscript of the article published in final edited form as:Pośpiech, E., Chen, Y., Kukla-Bartoszek, M., Breslin, K., Aliferi, A., Andersen, J. D., … Kayser, M. (2018). Towards broadening Forensic DNA Phenotyping beyond pigmentation: Improving the prediction of head hair shape from DNA. Forensic Science International: Genetics. https://doi.org/10.1016/j.fsigen.2018.08.017

Street, London, United Kingdom

[8]Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Frederik V's Vej 11, DK-2100 Copenhagen, Denmark

[9]Department of Genetic Identification, Erasmus MC University Medical Center Rotterdam, P.O. Box 2040, 3000 CA Rotterdam, Netherlands

[10]Institute of Legal Medicine, Medical Faculty, University of Cologne, Melatengürtel 60/62, D-50823 Cologne, Germany

[11]Forensic Genetics Unit, Institute of Forensic Sciences, R/ San Francisco s/n, Faculty of Medicine, 15782, University of Santiago de Compostela, Santiago de Compostela, Spain

[12]Division of Biological Traces, Netherlands Forensic Institute, P.O. Box 24044, 2490 AA, The Hague, The Netherlands

[13]Zurich Institute of Forensic Medicine, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

[14]Institute of Legal Medicine, Medical University of Innsbruck, Müllerstrasse 44, 6020 Innsbruck, Austria

[15]Department of Dermatology, Collegium Medicum of the Jagiellonian University, Skawińska st. 8, 31-066, Kraków, Poland

[16]Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah, KSA

[17]Forensic Science Program, The Pennsylvania State University, 13 Thomas Building, University Park, PA 16802, USA

[18]Institute of Legal Medicine, University of Münster, Röntgenstr. 23, 48149 Münster, Germany

[19]Chinese Academy of Sciences Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road Shanghai, 200031, P.R. China

[20]State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development,

School of Life Sciences, Fudan University, 2005 Song Hu Road Shanghai, 200438, P.R. China

[21]School of Life Science and Technology, Shanghai-Tech University, 393 Middle Huaxia Road, Pudong, Shanghai, 201210, P.R. China

[22]Queensland Institute of Medical Research, Brisbane, Australia

*Corresponding author: phone +31-10-7038073, e-mail m.kayser@erasmusmc.nl

Highlights

- Practical use of Forensic DNA Phenotyping is currently restricted to pigmentation traits
- Genetic knowledge underlying variation in human head hair shape has recently improved
- Hair shape prediction was evaluated here with an extended SNP set and independent test samples
- New hair shape prediction model was developed from European and Asian data
- New model provides improved prediction accuracy relative to previous models

Abstract

Human head hair shape, commonly classified as straight, wavy, curly or frizzy, is an attractive target for Forensic DNA Phenotyping and other applications of human appearance prediction from DNA such as in paleogenetics. The genetic knowledge underlying head hair shape variation was recently improved by the outcome of a series of genome-wide association and replication studies in a total of 26,964 subjects, highlighting 12 loci of which 8 were novel and introducing a prediction model for Europeans based on 14 SNPs. In the present study, we evaluated the capacity of DNA-based head hair shape prediction by investigating an extended

set of candidate SNP predictors and by using an independent set of samples for model validation. Prediction model building was carried out in 9,674 subjects (6,068 from Europe, 2,899 from Asia and 707 of admixed European and Asian ancestries), used previously, by considering a novel list of 90 candidate SNPs. For model validation, genotype and phenotype data were newly collected in 2,415 independent subjects (2,138 Europeans and 277 non-Europeans) by applying two targeted massively parallel sequencing platforms, Ion Torrent PGM and MiSeq, or the MassARRAY platform. A binomial model was developed to predict straight vs. non-straight hair based on 32 SNPs from 26 genetic loci we identified as significantly contributing to the model. This model achieved prediction accuracies, expressed as AUC, of 0.664 in Europeans and 0.789 in non-Europeans; the statistically significant difference was explained mostly by the effect of one *EDAR* SNP in non-Europeans. Considering sex and age, in addition to the SNPs, slightly and insignificantly increased the prediction accuracies (AUC of 0.680 and 0.800, respectively). Based on the sample size and candidate DNA markers investigated, this study provides the most robust, validated, and accurate statistical prediction models and SNP predictor marker sets currently available for predicting head hair shape from DNA, providing the next step towards broadening Forensic DNA Phenotyping beyond pigmentation traits.

Keywords: head hair; hair shape; externally visible characteristics; DNA prediction; Forensic DNA Phenotyping; targeted massively parallel sequencing

1. Introduction

Predictive DNA analysis of externally visible characteristics (EVCs), also referred to as Forensic DNA Phenotyping (FDP), is a fast growing area in forensic genetics. FDP uses DNA evidence to characterize unknown donors of crime scene traces who cannot be identified with

standard DNA profiling, to allow focussed investigation aiming to find them [1-2]. Outside the forensic field, DNA prediction is applied in anthropology and paleogenetics to reconstruct the appearance of deceased persons from (ancient) DNA analysis of (old) human remains [3-4]. However, EVCs for which both, statistical models providing reasonably high accuracies as well as validated genotyping methods reliably generating data from challenging DNA samples have been established, are currently restricted to the three pigmentation traits, eye, hair and skin colour [5-23].

The recent years have witnessed improvement in the genetic knowledge of human appearance for several other EVCs, most notably: body height [24-26], head hair loss in men [27-28], head hair shape [29-30], and facial shape [31-34]. For other EVCs, the first genes have been recently identified; these include facial hair and hair greying [29], and ear morphology [35-36]. Moreover, for two EVCs, DNA-based prediction models with improved accuracy relative to earlier models [37-39] were reported last year, i.e. for hair loss in men [28] and for head hair shape [30].

For human head hair shape, a series of genome-wide association studies (GWASs) and replication studies in a total of 26,964 subjects from Europe and other regions recently highlighted 12 genetic loci of which 8 were not previously known to be involved [30]. In this previous study, a prediction model based on 14 SNPs achieved prediction accuracy expressed as the area under the receiver operating characteristic curve (AUC), of 0.66 in 6,068 Europeans, and 0.64 in 977 independent Europeans [30]. These results represented improved accuracy relative to the first prediction model for hair shape that was based on 3 SNPs in 528 Europeans (AUC 0.622) reported previously by the EUROFORGEN-NoE Consortium [37].

In the present study of the EUROFORGEN-NoE Consortium and partners, head hair shape prediction capacity was further evaluated with data from 9,674 European and non-European donors used previously [30], and by considering 90 SNPs as candidate predictors in

the model building procedure. Moreover, an independent set of 2,415 Europeans and non-Europeans, for which genotype and phenotype data were newly collected, was used for subsequent model validation to derive prediction accuracy estimates. From this dataset and with suitable statistical methods, the best model for the prediction of head hair shape, achieving maximal accuracy with a minimal number of SNPs, was established and compared to previously reported models.

2. Materials and methods

2.1. Collection of samples for genotyping used for model validation

A total of 2,415 samples were collected by 10 participants of the EUROFORGEN-NoE consortium and one additional partner from the USA (Table 1) for genotyping of 90 hair shape candidate SNPs and for subsequent prediction model validation. All samples were obtained with informed consent. Samples were collected from random individuals by a medical doctor during clinical examination or by a researcher.

Phenotyping was performed through direct inspection by a dermatology specialist combined with the interview, by interview combined with the questionnaire or through evaluation of high quality photographs performed by independent researchers (Table 1). We made an effort to collect detailed information on hair shape status by applying a specific phenotypic regime, marking an advantage of the current research over our previous study [30]. If possible, samples were phenotyped according to a 6-point scale proposed by Eriksson et al. (Supplementary Text S5 Figure 1 in [40]), where 0 refers to stick straight, 1 – slightly wavy, 2 – wavy, 3 – big curls, 4 – small curls and 5 – very tight curls. In other cases, samples were categorized to 3 categories only including straight, wavy and curly hair. For the purpose of statistical calculations, a simplified classification was used for all the samples, where stick straight and straight accounted for ‘straight’ hair category, slightly wavy and wavy were

treated as the single category 'wavy' whereas big curls, small curls and very tight curls were incorporated into the category 'curly'. Wavy and curly categories were pooled for some analyses.

The majority of the study participants were of European ancestry (N=2,138), while the remaining individuals (N=277) were of non-European ancestry (Table 2). For a subset of samples with uncertain bio-geographic ancestry, the program ADMIXTURE [41] was used to infer the ancestry proportions of these individuals. Non-European samples consisted of samples assigned to one of the six following bio-geographic ancestries: European / non-European admixed (admixed-EUR), Africans (AFR), admixed Americans (AMR), Middle East, South Asians (SAS), and East Asians (EAS). This independent set of 2,415 samples was used as a model validation set.

2.2. Selection of SNPs used for model building

The study involved analysis of 90 candidate SNPs for human head hair shape, selected from various sources to minimize the chance of missing important markers for the final prediction model (Supplementary Table 1). The core 35 SNPs from a large GWAS and meta-analysis reported recently, were selected from the initial list of 706 SNPs with *P* values < $5x10^{-8}$ based on LD analysis where SNPs with independent effects were retained in each region [30] (Supplementary Figure 1). An additional set of 55 SNPs was chosen by a thorough literature search performed in parallel to the above-mentioned GWAS. Criteria to select SNPs from literature were: i) genes previously associated with hair shape (*TCHH, EDAR, WNT10A, FRAS1, OFCC1, TRAF2, PRSS53, PADI3, LOC105373470, S100A11, LCE3E, LOC391485*) [29,37,40,42-44]; ii) genes potentially involved in human hair morphogenesis and growth (*VDR*) [45-46]; iii) genes with expression patterns in the hair follicle (*IGFBP5*) [47]; iv) genes involved in protein-protein interactions reported to have a role in the hair structure

moulding (*RPTN*) [48]; v) genes previously associated with hair loss in men (*WNT10A*, *TARDBP*, *SUCNR1*, *EBF1*, *HDAC9*) [38,49], as a correlation between the genetics of hair morphology and hair loss has been recently suggested [49]; and vi) genes associated with pathological conditions of human hair structure (*KRT71*, *KRT74*, *LIPH*, *P2RY5*) [50-53]. In cases where no candidate SNPs within the selected genes have been suggested in the literature, SNP selection was performed using frequency information from The 1000 Genomes Project (http://grch37.ensembl.org/Homo_sapiens/Info/Index) [54]. In such cases, SNPs showing the highest allele frequency differences between EUR and AFR and/or between EUR and EAS, i.e. population groups with large differences in hair shape variation, were selected as the final candidates. The complete list of the 90 SNPs used is provided in Supplementary Table 1.

2.3. DNA extraction and multiplex SNP genotyping

Samples in the form of buccal swabs, saliva or whole blood were collected, DNA-extracted and SNP-genotyped by the participating laboratories (Table 1). Different methods for DNA extraction were used including: PrepFiler™ Forensic DNA Extraction Kit (ThermoFisher Scientific), QIAamp DNA Mini Kit (Qiagen), QIAamp DNA Blood Mini Kit (Qiagen), Chelex™ extraction, Gentra Puregene Buccal Cell Kit (Qiagen), Qiagen EZ1 DNA Investigator kit (Qiagen) and in-house salting-out method. Two targeted massively parallel sequencing platforms, that is Ion Torrent Personal Genome Machine (Ion Torrent PGM) (Thermo Fisher Scientific, Waltham, MA USA) and MiSeq (Illumina, San Diego, CA USA), as well as MassARRAY platform (Agena Bioscience) [55] were applied for multiplex genotyping of the 90 SNPs, depending on the participating laboratory (Table 1). Missing SNP genotypes were at the level of 3.8% and have been predicted using the mean substitution method implemented in R version 3.2.3.

2.3.1 Multiplex genotyping using Ion Torrent PGM

Samples were genotyped using Ion AmpliSeq™ Technology and the Ion Torrent Personal Genome Machine (PGM) system (Thermo Fisher Scientific). Primers for 96-SNP panel in one primer pool, including 90 candidate hair shape SNPs and 6 additional SNPs (IrisPlex SNPs: rs12913832, rs1800407, rs12896399, rs16891982, rs1393350, rs12203592 [5]) used for the purpose of technology validation, were designed with support of Thermo Fisher Scientific. DNA (1-5 ng) was amplified in 10 μL of PCR reaction and using one Ion AmpliSeq™ primer pool (2x). DNA libraries were prepared using the Ion AmpliSeq™ Library Kit 2.0, quantified with the Agilent High Sensitivity DNA Kit (Agilent Technologies) or Qubit dsDNA High-Sensitivity Assay Kit (Thermo Fisher Scientific), and normalized to 100 pM. DNA libraries for 32-40 samples were combined in equal ratios and subjected to template preparation with the Ion PGM HiQ OT2 Kit and the Ion OneTouch 2 System (Thermo Fisher Scientific). Alternatively, template preparation was conducted with Ion PGM™ IC 200 Kit using the Ion Chef System (Thermo Fisher Scientific). Templating was performed according to manufacturer's directions with the exception of the 100 pM DNA library volume that was increased from 2 to 5μL. Sequencing was conducted with the Ion PGM Hi-Q Sequencing Kit using Ion 318 Chips v2 or Ion PGM™ IC 200 Kit (Thermo Fisher Scientific), 200 bp read chemistry and 520 flows per run. Raw data were analysed using Torrent Server v5.0.5 and DNA SNP alleles were called using variantCaller v5.0.4.0 (Thermo Fisher Scientific). For quality assurance, genotypes of 12 out of the 96 SNPs analysed (rs17646946, rs11803731, rs4845418, rs12130862, rs7349332, rs1268789, rs5919324, rs1998076, rs929626, rs12565727, rs756853, and rs4679955) were compared with genotype data previously obtained with an alternative technology, that is primer extension minisequencing based on

SNaPshot chemistry (Thermo Fisher Scientific) in 491 overlapping samples [37]. All 12x491 genotypes were consistent between these two technologies.

2.3.2 Multiplex genotyping using Illumina MiSeq

Samples were genotyped using Illuminas MiSeq Reagent Kit v2 Technology and MiSeq FGx System in RUO mode (Illumina). Primers (including adapter sequences) for 91 SNPs in 3 separate primer pools (90 candidate hair shape SNPs and 1 additional SNP, that is rs12913832 used for control purposes) were designed in house. DNA (2 - 20 ng/sample) was amplified in 10 or 20 µL PCR reactions using three separate reactions and the Kapa Multiplex PCR Mix (Kapa Biosystems). PCR products from the three reactions were cleaned (using an in house bead protocol). The cleaned products were then quantified using Agilent High Sensitivity DNA Kit (Agilent Technologies) or Qubit dsDNA High-Sensitivity Assay Kit (Thermo Fisher Scientific), normalized and pooled into a new plate where they were indexed. Additional separate steps were performed by KCL, these include library preparation performed using the KAPA Hyper Prep kit for Illumina platforms (Roche) at half volume with 9 library amplification cycles, Illumina® TruSeq™ adapters were added during the library preparation stage and libraries were quantified with the Kapa Library Quantification kit for Illumina platforms (Roche). Finally, a DNA library of all individuals was pooled, normalized to 2 nM (IUPUI) / 4 nM (KCL) and sequenced using the MiSeq 300-cycle v2 reagent kit (Illumina). Raw data was aligned to a custom-made fasta file containing the reference sequences for all amplicons using the mem algorithm within BWA (http://bio-bwa.sourceforge.net/) specifying a minimum seed length of 40bp [56]. The sequence alignment/map (SAM) file was converted using SAMtools into a BAM file [57]. Variant calling was made with GATK's Unified Genotyper (v 2.8-1) using default parameters with no down-sampling [58]. Alternatively, data were analyzed using FDSTools software (https://pypi.python.org/pypi/fdstools/).

2.3.3 Multipex genotyping using MassARRAY® platform

Genotyping of 84 SNPs was conducted by the CEGEN-PRB2 USC node using the iPlex® Gold chemistry and MassARRAY platform, according to manufacturer's instructions (Agena Bioscience, San Diego, EEUU). Six SNPs were excluded due to technical problems (rs4480966, rs78544048, rs72696940, rs11575161, rs11204925, rs6658216). Genotyping assays were designed using the Agena Bioscience MassARRAY Assay Designer 4.1 software. 84 SNPs were genotyped in 4 assays. PCR reactions were set up in a 5 µl volume and contained 20 ng of template DNA, 1× PCR buffer, 2 mM MgCl2, 500 µM dNTPs and 1 U/reaction of PCR enzyme. A pool of PCR primers was made at a final concentration of each primer of 100 nM (IDT, Integrated DNA technologies, Newark, EEUU). The thermal cycling conditions for the reaction consisted of an initial denaturation step at 95°C for 2 minutes, followed by 45 cycles of 95°C for 30 seconds, 56°C for 30 seconds and 72°C for 1 minute, followed by a final extension step of 72°C for 5 minutes. PCR products were treated with 1.7 U shrimp alkaline phosphatase by incubation at 37°C for 40 min, followed by enzyme inactivation by heating at 85°C for 5 min to neutralize unincorporated dNTPs. The iPLEX GOLD reactions were set up in a final 9 µl volume and contained 0.222x iPLEX buffer Plus, 0.222x iPLEX Termination mix and 1.35 U/reaction iPLEX enzyme. An extension primer mix was made to give a final concentration of each primer between 0.52 µM and 1.57 µM (IDT, Integrated DNA technologies, Newark, EEUU). The thermal cycling conditions for the reaction included an initial denaturation step at 95°C for 30 seconds, followed by 40 cycles of 95°C for 5 seconds, with an internal 5 cycles loop at 52°C for 5 seconds and 80°C for 5 seconds, followed by a final extension step of 72°C for 3 minutes. The next step is to desalt the iPLEX Gold reaction products with Clean Resin following the manufacturer's protocol. The desalted products were dispensed onto a 384 Spectrochip II using an RS1000

Nanodispenser and spectra were acquired using the MA4 mass spectrometer, followed by manual inspection of spectra. Samples were genotyped analysed with Typer 4.0.163 software using standard SNP genotyping parameters. All assays were performed in 384-well plates, including negative controls and a trio of Coriell samples (NA10860, NA10861 and NA11984) for quality control. 10% random samples were tested in duplicate and the reproducibility was 100%.

2.4 Multinomial and binomial logistic regression analyses

The multinomial logistic regression modelling specifications were as following. Consider hair shape, y, to be three categories straight, wavy, and curly, which are determined by the genotype, *x*, of *k* SNPs, where *x* represents the number of minor alleles per *k* SNP. Let $\pi_1, \pi_2,$and π_3 denote the probability of categories straight, wavy, and curly respectively. The multinomial logistic regression can be written as

$$\text{logit}(\Pr(\text{y} = \text{straight}|x_1 \dots x_k)) = \ln\left(\frac{\pi_1}{\pi_3}\right) = \alpha_1 + \sum_{i=1}^{k} \beta(\pi_1)_i x_i$$

$$\text{logit}(\Pr(\text{y} = \text{wavy}|x_1 \dots x_k)) = \ln\left(\frac{\pi_2}{\pi_3}\right) = \alpha_2 + \sum_{i=1}^{k} \beta(\pi_2)_i x_i$$

where α and β can be derived in the training set. Hair shape of each individual in the testing set can be probabilistically predicted based on his or her genotypes and the derived αand β,

$$\pi_1 = \frac{\exp(\alpha_1 + \sum_{i=1}^{k} \beta(\pi_1)_i x_i)}{1 + \exp\left(\alpha_1 + \sum_{i=1}^{k} \beta(\pi_1)_i x_i\right) + \exp\left(\alpha_2 + \sum_{i=1}^{k} \beta(\pi_2)_i x_i\right)}$$

$$\pi_2 = \frac{\exp\left(\alpha_2 + \sum_{i=1}^{k} \beta(\pi_2)_i x_i\right)}{1 + \exp\left(\alpha_1 + \sum_{i=1}^{k} \beta(\pi_1)_i x_i\right) + \exp\left(\alpha_2 + \sum_{i=1}^{k} \beta(\pi_2)_i x_i\right)}$$

$$\pi_3 = 1 - \pi_1 - \pi_2$$

Categorically, the shape category with the $\max(\pi_1, \pi_2, \pi_3)$ was considered as the predicted shape type.

The binomial logistic regression modelling specifications were as following. Consider hair shape, y, to be two categories straight and non-straight, which are determined by the genotype, *x*, of *k* SNPs, where *x* represents the number of minor alleles per *k* SNP. Let p denote the probability of straight, and $1 - \mathrm{p}$ is the probability of non-straight. The binomial logistic regression can be written as

$$\mathrm{logit}(\Pr(\mathrm{y} = \mathrm{straight}|x_1 \dots x_k)) = \ln\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = \alpha + \sum_{i=1}^{k} \beta_i x_i$$

where α and β can be derived in the training set. Hair shape of each individual in the testing set can be probabilistically predicted based on his or her genotypes and the derived αand β,

$$\mathrm{p} = \frac{\exp(\alpha + \sum_{i=1}^{k} \beta_i x_i)}{1 + \exp\left(\alpha + \sum_{i=1}^{k} \beta_i x_i\right)}$$

Categorically, the shape category with the $\max(\mathrm{p}, 1 - \mathrm{p})$ was considered as the predicted shape type.

2.5. Prediction marker selection, prediction model building and validation

Figure 1 shows the study design including the sample sets used. Starting with candidate SNPs from the literature, we additionally considered the outcomes from a GWAS meta-analysis previously conducted in three European cohorts including QIMR (Queensland Institute of Medical Research study), RS (Rotterdam Study), and TwinsUK referred to as META:Discovery in our previous publication [30], which considered a total of 16,763 subjects. From the list of 90 candidate SNPs, 60 (66.7%) showed nominal-significant association (*P* value < 0.05) in this dataset, of which 45 (75%) displayed genome-wide significant association (*P* value < $5\text{x}10^{-8}$). From a set of 60 SNPs showing nominal-significant association, a minimal set of SNP predictors for maximizing the prediction accuracy was

chosen using step-wise regression according to the Akaike information criterion (AIC) in a total of 9,674 subjects. This included 6,068 Europeans from QIMR, 2,899 Chinese from TZL (Chinese Taizhou Longitudinal) and 707 Xinjiang Uyghurs known to be of 50% European and 50% East Asian admixed ancestry [59-60] from our previous study [30], where further details on these cohorts can be found. Selection of SNP predictors was performed on hair shape that was classified into 2 categories: straight vs. non-straight, using binomial logistic regression (BLR). Additionally, 3-category classification as straight vs. wavy vs. curly has been applied for marker selection using multinomial logistic regression (MLR). Final prediction models were built using data from QIMR, TZL and UYG studies (model building set).

The developed prediction models were verified for their accuracy using a dataset of 2,415 independent samples described in section 2.1; this dataset was used as a model validation set. Prediction accuracy was evaluated by estimating several parameters including AUC, sensitivity, specificity, negative prediction value (NPV) and positive prediction value (PPV). Sensitivities and specificities, PPVs and NPVs were calculated using confusion matrices considering the predicted probability > 0.5 as the predicted hair shape type. Prediction accuracy parameters were estimated for all samples included in the model validation set, considering all samples together as well as for the European and non-European samples separately. Statistical analyses and result visualization were conducted in R version 3.2.3.

2.6. Significance testing

To test if different logistic models consisting of different marker sets may provide statistically different prediction results, we used the student's t-test to compare the absolute values of the residuals of the predicted probabilities of having straight hair. The Model 1 consisted of 3

SNPs [37], the Model 2 consisted of 14 SNPs [30], and the Model 3 consisted of 32 SNPs from the current study. We trained Model 1 and Model 3 in 6068 European subjects from QIMR together with 3606 Asian subjects from UYG and TZL, whereas Model 2 was trained in 6068 European subjects from QIMR. We then applied these models to 2415 EUROFORGEN-NoE subjects and obtained the predicted probabilities of having straight hair for all the tested subjects.

To test if there is a statistical difference in the percentage of variation in head hair shape explained by *EDAR* and by *TCHH*, we derived the Nagelkerke's pseudo R^2 from logistic models (straight vs. non-straight) separately from *EDAR* and *TCHH*, in the combined samples of TZL, UYG, and QIMR. The ANOVA was used to test if the R^2 values from different models are statistically different. All statistical analyses were conducted in R version 3.3.1 unless otherwise specified.

3. Results and Discussion

3.1 Selection of SNP predictors and prediction model building

3.1.1 BLR model building for predicting straight vs. non-straight hair

The BLR analysis identified 32 significantly contributing SNPs from 26 genetic loci given a 2-category classification (straight vs. non-straight hair). In our previous study [30], 14 SNPs from 14 genes (13 SNPs showing genome-wide significant association in the META:Discovery) were used for prediction modelling. Eleven of them were included in the current BLR model. The remaining 3 SNPs were replaced with substitute LD SNPs due to technical difficulties (Table 3). The final BLR model further included additional 9 SNPs selected from our previous GWA study [30] and 9 SNPs selected from the literature (see Table 3). Five of those have been directly associated with hair morphology in several previous studies, including rs3827760 in *EDAR*, rs11803731 in *TCHH*, rs4672907 in

LOC391485, rs3007671 in *NBPF18P/S100A11* and rs436034 in *FRAS1* [29,37,40,42-43,59,61]. SNP rs3827760 in *EDAR* was not tested for association in the previous META:Discovery as this SNP is almost monomorphic in Europeans and therefore did not pass the quality control for Europeans in the GWAS analysis [30]. However, due to its known impact on hair morphology determination in East Asians [43,59,61], and its recently reported association in admixed Latin Americans [29], this SNP was also included in the prediction modelling. As evident from Table 3, this SNP ranks at the first position in the BLR analysis due to the use of East Asians (TZL, UYG) in addition to Europeans (QIMR), in the marker selection dataset. While *EDAR* is considered to be the most important genetic determinant of straight hair in East Asians [43,59,61], *TCHH* is believed to be a major straight hair gene in Europeans [37,40,42]. The SNP rs11803731 *TCHH* is suggested to be the most likely functional variant [42] and ranks at 2nd place in the current BLR analysis (Table 3).

Notably, the step-wise regression analysis leading to the final BLR model also highlighted 4 SNPs from 3 new candidate genes, *RPTN*, *KRT71* and *LIPH* [40,51,53]. *RPTN*, similar to *TCHH*, is a member of the 'fused' gene family that is localized in the epidermal differentiation complex (EDC) on chromosome 1. Proteins encoded by the *RPTN* and *TCHH* genes are implicated in strong interaction at the biological level during formation of cornified cell envelope (CE) known to play a crucial role in mechanical protection of the hair follicle [40,48,62]. Recently, interaction between *RPTN* and *TCHH* has been confirmed at the statistical level and shown to facilitate straight hair formation in Europeans [63]. The SNP rs3001978 in *RPTN* achieved significance at the level of *P* value=1.00x10^{-8} in the previous META:Discovery GWAS [30] and is ranked 26th in the current BLR model. Genes *KRT71* and *LIPH* were previously associated with pathological hair structure like woolly hair (WH, characterized by abnormally tightly curled hair) [51,53]. Possible involvement of genes responsible for abnormal hair structure in the determination of natural variation of hair

morphology has been suggested previously [52,63]. Two SNPs from the *KRT71* gene, rs10783518 (10^{th}) and rs585583 (16^{th}) showed significant association at *P* value = $2.67x10^{-4}$ and *P* value = $2.65x10^{-3}$ in the previous META:Discovery GWAS analysis [30].

SNPs in the novel genes selected from the literature have been chosen using frequency information from The 1000 Genomes Project. Therefore, it should be pointed out that the selected polymorphisms may not have functional relevance. Future studies, including fine mapping of the novel genes and some functional analyses may reveal whether particular DNA variants are causal or are just in LD with functional SNPs. However, regardless to the functional meaning of the variants, they can still provide information on the phenotype through LD phenomenon and be useful in prediction modelling, as demonstrated here.

3.1.2. MLR model building for predicting straight vs. wavy vs. curly hair

Previous studies have evaluated head hair shape predictability at the level of straight vs. non-straight hair only [30,37]. In the present study, we additionally took a step forward and assessed hair shape predictability considering a 3-category classification (straight vs. wavy vs. curly hair) using MLR. The 3-category classification MLR approach highlighted 33 SNPs from 29 genetic loci, of which 27 overlapped with the 32 SNPs identified in the BLR model (Table 3). Similar to the BLR analysis, the first two positions in the MLR-based ranking approach were rs3827760 in *EDAR* and rs11803731 in *TCHH*, as may be expected. The rank 3 SNP, however, was rs11150606 in *PRSS53*, which was not selected in the BLR analysis. *PRSS53* encodes Protease Serine S1 family member 53. This SNP was identified by a GWAS in admixed Latin Americans [29], and replicated in our recent GWA study at a nominal significance level [30]. In addition to rs11150606, there were 5 more SNPs in the MLR model that were not included in the BLR model (Table 3).

3.2 Validating the prediction models and estimating prediction accuracy for head hair shape

The BLR and MLR models built in the 9,674 Europeans, Asians, and admixed European-Asians were subjected to model validation using an independent set of 2,415 European and non-European samples that were not previously used for model building and also not for prior marker discovery. It is generally recommended in genetic prediction studies to use different datasets for marker discovery, model building, and model validation, respectively, which we achieved with the current study [64-65].

3.2.1 BLR model validation for predicting straight vs. non-straight hair

The binomial model based on 32 SNPs was found to predict straight vs. non-straight hair with AUC=0.679 in the entire model validation set (N=2,415). The achieved sensitivity was 0.840, which means that out of 946 individuals of straight hair 795 were predicted correctly as being straight haired. Considerably lower level of prediction specificity compared to sensitivity was achieved at 0.364 which shows a reduced ability of the model to detect curly/wavy hair; out of 1469 individuals of wavy/curly hair 935 of them were incorrectly classified as straight haired. The positive predictive (PPV) and the negative predictive (NPV) values were 0.460 and 0.780, respectively. PPV value of 0.460 means that in all cases in which hair shape was classified as straight, for 46% of individuals prediction result was correct. NPV value at the level of 0.780 means that out of all non-straight hair classifications in 78% of them prediction was correct and individuals indeed had non-straight hair.

When performing prediction model validation by considering bio-geographic ancestry of the samples used, we obtained an AUC value for the European sub-sample set (N=2,138) at 0.664, and a statistically significantly (*P* value = 0.02) higher AUC value for non-Europeans (N=277) at 0.789 (Table 4). Sensitivity was rather similar for the European and non-European sample sets. Out of 870 Europeans with straight hair, 732 (sensitivity of 0.841) were correctly

predicted as straight haired, and 63 out of the 76 straight-haired non-Europeans (0.829) were correctly predicted as such. In contrast, specificity was considerably higher in the non-Europeans relative to the Europeans (Table 4). Out of the 1,268 wavy or curly haired Europeans and the 201 non-Europeans, 434 (specificity of 0.342) and 100 (0.498) were correctly predicted as non-straight haired (Table 4).

The increase of prediction accuracy estimates in the non-Europeans, which is driven by an increased specificity of straight hair prediction, is likely caused by the effect of the *EDAR* SNP in the BLR model. Aiming to get more insights, we conducted prediction analyses in non-Europeans at the level of continental groups. Although these results should be taken with considerable caution because of the low number of samples per ancestry group, they suggest that the improved prediction accuracy obtained in the non-Europeans originates mostly from the East Asians (EAS) (AUC=0.694), Admixed Americans (AMR) (AUC=0.750) and the Middle Easterners (AUC=0.833) for which AUCs were increased. In contrast, the AUC values were considerably lower for all other ancestry groups (Supplementary Table 2). Notably, AUC values for EAS, AMR and Middle East decreased substantially when the *EDAR* SNP rs3827760 was excluded from the model (0.581, 0.687 and 0.589, respectively) (Supplementary Table 2), demonstrating the *EDAR* SNP effect on hair shape prediction in these ancestry groups. The *EDAR* gene encodes a member of the tumour necrosis factor receptor family and was associated with several phenotypes of epidermal appendages including ectodermal dysplasia, dental morphology and sweat gland density [66-67]. *EDAR* rs3827760, used in the prediction model, is a coding SNP showing one of the strongest signals of natural selection in human genomes. This SNP has been previously associated with thick and straight hair in East Asians [43,59,61]. The derived G allele, which drives the noted association, is observed with a high frequency in East Asians and Native

Americans, while in Europeans and Africans rs3827760 is almost monomorphic for the ancestral A allele [43].

These as well as previous findings [42-43,59,61] suggest different mechanisms of straight hair in Europeans and Asians, at least with regard to the *EDAR* effect. The odds ratio for *EDAR* rs3827760 in East Asians was previously reported at the level of about 2-2.5 in two independent studies, explaining 2-4% of the total variation in hair structure in these datasets [43,59]. Therefore, the size of the *EDAR* gene effect in East Asians seems to correspond with the effect of *TCHH* in Europeans, although a stronger effect of *EDAR* compared to *TCHH* was recently suggested by analysis of Uyghur samples, a population suggested be of 50% European and 50% East Asian admixed ancestry [59]. When looking at the samples used for prediction model building in the current study, rs3827760 in *EDAR* explains 16.2% (Nagelkerke R^2) of the variation in hair shape in TZL, UYG and QIMR, while significantly lower (*P* value = 7.02×10^{-293}) variation is explained by *TCHH* at 0.3%. Although the *EDAR* gene seems to be a major hair morphology gene in East Asians, it does not explain the full heritability of hair shape variation in this ethnic group, confirming that the biology of Asian hair shape is more complex. Although improvement of hair shape prediction after *EDAR* inclusion was expected in the AMR group (with 40% frequency of the G allele), given their partial ancestry from East Asia based on their initial migration history, we also observed an increased AUC value in Middle East, where only 3% of individuals were found to carry the G allele. However, in this ancestry group, wavy and curly hair exists almost exclusively, which may be associated with the lack of the G allele in *EDAR* rs3827760. Additional studies are needed to evaluate the role of *EDAR* in different non-European ancestry groups.

When the distribution of the obtained non-straight hair probabilities was analysed, the results were in general agreement with general knowledge on the global distribution of hair shape variation i.e., decreasing prevalence of wavy and curly hair from Africans through

Europeans towards Asians (Supplementary Figure 2). The highest probability values for non-straight hair were seen for Africans and the lowest for East Asians with Europeans, South Asians and Native Americans in between. Moreover, the degree of variation in probability values was the highest in Europeans (Var=0.024) and Africans (Var=0.019) and lowest in East Asians (Var=0.014) (Supplementary Figure 2). This outcome is concordant with the results obtained in our previous study, where 2,504 worldwide subjects from the 1000-Genomes Project panel were analysed using the previous 14-SNP model [30].

As illustrated in Figure 2 and Supplementary Table 3, the individual contributions of the 32 SNPs used in the BLR model towards the overall prediction accuracy substantially differed between the markers. A strong effect of the *EDAR* gene in non-Europeans, but weak on Europeans, was noted. In particular, the single *EDAR* variant (rs3827760) provides a high degree of accuracy for straight vs. non-straight hair prediction with AUC=0.742 in non-Europeans, but only 0.505 in Europeans. The highest contribution to the accuracy of hair structure prediction in Europeans came from three SNPs that are rs11803731 in *TCHH* (AUC change 0.089), rs1268789 in *FRAS1* (AUC change 0.02) and rs80293268 in *ERRFI1/SLC45A1* (AUC change 0.016). Besides the high contribution of the *EDAR* SNP, noticeable impact on the AUC value (AUC increase > 0.005) was provided by rs11803731 in *TCHH*, rs1268789 in *FRAS1*, rs310642 in *PTK6* and rs2219783 in *LGR4*, respectively, in non-Europeans. Among loci with the largest impact on AUC, three of them *ERRFI1/SLC45A1*, *PTK6* and *LGR4* were newly discovered in our recent GWAS [30]. The input of the remaining markers was found to be considerably lower (Figure 2 and Supplementary Table 3).

3.2.2 Impact of age and sex on straight vs. non-straight hair prediction

In a previous study conducted on >1,600 individuals of European ancestry, males were found to be ~5% more likely to have straight hair than females, and additionally, curliness of hair in males was reported to increase slightly with age [42]. In contrast, no significant effect from age and sex on hair morphology was found in previous work performed by the EUROFORGEN-NoE Consortium, but the number of samples analysed was considerably smaller (N=528) [37]. We therefore evaluated age and sex as additional factors in the BLR modelling. As illustrated in Table 4, sex and age had positive but slight effects on straight vs. non-straight hair prediction accuracy. AUC increased, but statistically insignificantly, from 0.679 to 0.695 (*P* value = 0.29) in the total model validation set, from 0.664 to 0.680 (*P* value = 0.36) in the European and from 0.789 to 0.800 (*P* value = 0.58) in the non-European subsets when considering age and sex. When testing for the impact of age and sex separately, both factors had similarly small effects (Supplementary Table 4). Notably, sex is typically available in forensic analyses due to the inclusion of amelogenin in standard DNA profiling, while age can be estimated via epigenetic profiling [68].

3.2.3 Comparison with previous models for straight vs. non-straight hair prediction

The BLR model introduced here provided improved prediction accuracy when comparing to all model previously proposed for head shape. The first reported model involved just 3 SNP predictors, rs11803731 in *TCHH*, rs1268789 in *FRAS1*, and rs7349332 in *WNT10A* based on previous work by the EUROFORGEN-NoE Consortium [37]. In that study, 528 samples from Poland were analysed and used for prediction modelling, achieving a cross-validated AUC at the level of 0.622 for straight vs. non-straight hair when using a logistic regression method. When these three SNP predictors were used for model building with the current model building dataset (N=9,674), a slightly lower AUC value for the European model validation subset (N=2,138) was achieved at 0.605. In the current study, a much larger sample size for

model building and an independent sample sets for model validation were used, in contrast to the previous study where the same dataset was used for model building and cross-validation. Therefore, the noted discrepancy either indicates a previous overestimation due to not using an independent validation set, and/or a higher phenotyping accuracy in the previously used samples. Most importantly, the accuracy of the 32-SNP BLR model achieved here for the total model validation set (AUC=0.679) is statistically significantly higher compared to the AUC we obtained when only 3 SNPs were analysed (AUC=0.605) (*P* value = 6.15×10^{-17}). Moreover, it also is statistically significantly higher compared to the recently proposed 14-SNP model (AUC=0.66) (*P* value = 1.12×10^{-5}) [30]. For the European subset, the accuracy increase provided by the new 32-SNP model relative to the previous 14-SNP model is rather small and statistically insignificant (0.664 versus 0.66), and further increased slightly when considering age and sex (0.68). However, a considerably stronger increase in prediction accuracy available with the new 32-SNP BLR model is seen for non-Europeans, for which an AUC of 0.789 (0.80 with sex and age) was achieved. When using only 3 SNPs previously applied by Pośpiech et al., (2015) [37] in the current model building set, to the current non-European model validation subset (N=277), a considerable and statistically significantly lower AUC (0.594 versus 0.789) was obtained (*P* value = 2.46×10^{-10}).

3.2.4 MLR model validation for predicting straight vs. wavy vs. curly hair

The multinomial regression model based on 33 SNPs (Table 3) predicted hair morphology in Europeans with AUC=0.666 for straight, 0.596 for wavy and 0.600 for curly hair in the model validation subset (Table 5). Although high sensitivity was achieved for straight hair (0.862), the reduced AUC for this category is caused by the lower level of specificity (0.324), as also seen in the BLR model. In particular, of the 1,268 curly and wavy haired individuals, 1,150 (90.7%) were misclassified as being straight-haired. Improved prediction accuracy was noted

for the non-European model validation subset in all three categories: straight (AUC=0.801), wavy (AUC=0.609) and curly hair (AUC=0.736). While a similar level of straight hair prediction sensitivity was reported for Europeans and non-Europeans (0.842), increased specificity was seen for non-Europeans (0.438) compared to Europeans (0.324). This result is correlated with increased, but still low, sensitivity for wavy and curly hair prediction at the level of 0.016 and 0.013, for EUR and non-EUR respectively. These results suggest the need for further studies to identify additional predictors needed to achieve more accurate predictions with higher phenotyping resolution. Moreover, including more head hair shape categories at the level of statistical calculations requires larger number of samples used, particularly for the testing purposes.

When testing the contribution of individual SNPs to the 3-category prediction accuracy, high impact, as expected, was noted for *EDAR* rs3827760 in non-Europeans and *TCHH* rs11803731 in Europeans (Supplementary Table 5). *EDAR* rs3827760 when used alone in the modelling, did not impact the prediction accuracy for Europeans with AUC at the level of ~0.5 for all three hair shape categories, which literally means random prediction. In contrast, this *EDAR* SNP provided high prediction accuracy for straight hair (AUC=0.742) and somewhat lower for curly (AUC=0.659) and wavy hair (AUC=0.565) in non-Europeans. *TCHH* rs11803731 used alone, provided more accurate prediction of hair shape in Europeans (AUC=0.593 for straight, 0.567 for wavy and 0.554 for curly hair) and also improved prediction accuracy in non-Europeans (AUC=0.769 for straight, 0.567 for wavy and 0.683 for curly hair), which is most probably explained by the effect of *TCHH* in admixed-EUR and AMR. Similar to the BLR model, subsequent SNPs with the highest impact on hair shape prediction in the MLR model in Europeans were rs1268789 in *FRAS1*, rs80293268 in *ERRFI1/SLC45A1*, and rs310642 in *PTK6*. These SNPs were also found to have a noticeable impact on the 3-category hair shape prediction in non-Europeans along with two additional

polymorphisms, that are *LGR4* rs2219783 and *PRSS53* rs11150606 affecting wavy and curly hair shape prediction in non-Europeans only (Supplementary Table 5). Sex and age had a positive impact on MLR prediction accuracy of some hair shape categories with AUC increase at ~0.01-0.02 (Supplementary Table 6).

Conclusions

Recent progress in better understanding the complex genetic architecture of some EVCs, head hair shape included, and the discovery of strongly associated SNPs, provides suitable resources for broadening Forensic DNA Phenotyping and other applications of DNA-based appearance prediction such as in anthropology and paleogenetics, beyond the currently practised pigmentation traits. Here, we developed a BLR model for straight vs. non-straight hair prediction based on 32 SNPs, and – for the first time - a MLR model for straight vs. wavy vs. curly hair prediction based on 33 SNPs, which largely overlap between the models. The new 32-SNP BRL model significantly improves 2-category hair shape prediction compared to the previously reported 3-SNP and 14-SNP models. Although non-Europeans revealed more accurate prediction outcomes than European, for both the BRL and the MRL models, the limited non-European data available here require additional studies involving larger numbers of samples from various human populations to further evaluate the model accuracy. However, the 32 SNPs used in the new BLR model only explain 12.1% of hair shape variation in the current model building set of 6,068 samples and, as evident from PPV and NPV values, the prediction outcome was correct in 46% of straight hair classifications and 78% of non-straight hair classifications. Thus, despite our ability to provide an improved model and DNA marker set for hair shape prediction, providing the next step towards broadening Forensic DNA Phenotyping beyond pigmentation traits, this study also demonstrates that the search for more

hair shape associated DNA variants and the investigation of their predictive value in independent samples needs to continue.

Acknowledgements

The authors thank all sample donors for their contribution to this project, Demi Wiskerke (Den Haag) for NFI sample collection, Christina Strobl and Bettina Zimmermann (Innsbruck) as well as Corinne Moser (Zurich) for valuable technical assistance. This research was supported by grants from the European Union Seventh Framework Programme no. 285487 (EUROFORGEN-NoE), the resources for Polish science in years 2012-2016 (no 20/7.PR/2012), and the National Science Centre in Poland no 2014/15/D/NZ8/00282. MK, LC and FL were supported by Erasmus MC. FL received additional support by the Erasmus University Rotterdam (EUR) Fellowship, the Chinese Recruiting Program 'The National Thousand Young Talents Award', National Key R&D Program of China (2017YFC0803501) and National Natural Science Foundation of China (91651507). SX acknowledges financial support from the Chinese Academy of Sciences (XDB13040100 and QYZDJ-SSW-SYS009) and the National Natural Science Foundation of China (NSFC) grant (91731303, 31771388, 31525014 and 31711530221). SW was supported by the US National Institute of Justice (NIJ) Grant 2014-DN-BX-K031, and the US Department of Defense (DOD) DURIP-66843LSRIP-2015. EP was supported by Foundation for Polish Science within the programme START 2017. LGS was supported by a training program of the Ministry of Economy and Competitiveness, Spain (as part of the Plan Estatal de Investigación Científica y Técnica y de Innovación 2013-20). AFA was supported by a postdoctorate grant funded by the Xunta de Galicia, Spain, as part of the Plan Galego de Investigación, Innovación e Crecemento 2011–2015, Axudas de apoio á etapa de formación postdoutoral, Plan I2C.

References

1. M. Kayser, P. de Knijff, Improving human forensics through advances in genetics, genomics and molecular biology, Nat. Rev. Genet. 12 (2011) 179–192.
2. M. Kayser, Forensic DNA phenotyping: predicting human appearance from crime scene material for investigative purposes, Forensic Sci. Int. Genet. 18 (2015) 33-48.
3. T.E. King, G.G. Fortes, P. Balaresque, M.G. Thomas, D. Balding et al., Identification of the remains of King Richard III. Nat Commun. 5 (2014) 5631.
4. I. Olalde, M.E. Allentoft, F. Sanchez-Quinto, G. Santpere, C.W.K. Chiang et al., Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. Nature 507 (2014) 225-228.
5. F. Liu, K. van Duijn, J.R. Vingerling, A. Hofman, A.G. Uitterlinden, et al., Eye color and the prediction of complex phenotypes from genotypes, Curr. Biol. 19 (2009) R192–R193.
6. W. Branicki, U. Brudnik, A. Wojas-Pelc, Interactions between HERC2, OCA2 and MC1R may influence human pigmentation phenotype, Ann. Hum. Genet. 73 (2009) 160–170.
7. J. Mengel-From, C. Børsting, J.J. Sanchez, H. Eiberg, N. Morling, Human eye colour and HERC2, OCA2 and MATP, Forensic. Sci. Int. Genet. 4 (2010) 323–328.
8. R.K. Valenzuela, M.S. Henderson, M.H. Walsh, N.A. Garrison, J.T. Kelch, et al., Predicting phenotype from genotype: normal pigmentation, J. Forensic Sci. 55 (2010) 315–322.
9. W. Branicki, F. Liu, K. van Duijn, J. Draus-Barini, E. Pospiech, et al., Model-based prediction of human hair color using DNA variants, Hum. Genet. 129 (2011) 443–454.

10. S. Walsh, F. Liu, K.N. Ballantyne, M. van Oven, O. Lao, et al., IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information, Forensic Sci. Int. Genet. 5 (2011) 170–180.
11. S. Walsh, A. Lindenbergh, S. Zuniga, T. Sijen, P. de Knijff et al., Developmental validation of the IrisPlex system: determination of blue and brown iris colour for forensic intelligence. Forensic Sci. Int. Genet. 5 (2011) 464 – 471.
12. O. Spichenok, Z.M. Budimlija, A.A. Mitchell, A. Jenny, L. Kovacevic, et al., Prediction of eye and skin color in diverse populations using seven SNPs, Forensic. Sci. Int. Genet. 5 (2011) 472–478.
13. S. Walsh, A. Wollstein, F. Liu, U. Chakravarthy, M. Rahu et al., DNA-based eye colour prediction across Europe with the IrisPlex system. Forensic Sci. Int. Genet. 6 (2012) 330-340.
14. E. Pośpiech, J. Draus-Barini, T. Kupiec, A. Wojas-Pelc, W. Branicki, Prediction of eye color from genetic data using Bayesian approach, J. Forensic Sci. 57 (2012) 880–886.
15. S. Walsh, F. Liu, A. Wollstein, L. Kovatsi, A. Ralf, et al., The HIrisplex System for simultaneous prediction of hair and eye colour categories including hair colour shade from DNA, Forensic Sci. Int. Genet. 7 (2013) 98–115.
16. Y. Ruiz, C. Phillips, A. Gomez-Tato, J. Alvarez-Dios, M. de Cal Casares, et al., Further development of forensic eye color predictive tests, Forensic Sci. Int. Genet. 7 (2013) 28–40.
17. V. Kastelic, E. Pośpiech, J. Draus-Barini, W. Branicki, K. Drobnic, Prediction of eye color in the Slovenian population using the IrisPlex SNPs, Croat. Med. J. 28 (2013) 381–386.

18. J.S. Allwood, S. Harbison, SNP model development for the prediction of eye colour in New Zealand, Forensic Sci. Int. Genet. 7 (2013) 444–452.

19. S. Walsh, L. Chaitanya, L. Clarisse, L. Wirken, J. Draus-Barini et al., Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage. Forensic Sci Int Genet. 9 (2014) 150-161.

20. E. Pośpiech, A. Wojas-Pelc, S. Walsh, F. Liu, H. Maeda et al., The common occurrence of epistasis in the determination of human pigmentation and its impact on DNA-based pigmentation phenotype prediction, Forensic Sci. Int. Genet. 11 (2014) 64–72.

21. O. Maroñas, C. Phillips, J. Söchtig, A. Gomez-Tato, R. Cruz et al., Development of a forensic skin colour predictive test. Forensic Sci. Int. Genet. 13 (2014) 34-44.

22. J. Söchtig, C. Phillips, O. Maroñas, A. Gómez-Tato, R. Cruz et al., Exploration of SNP variants affecting hair colour prediction in Europeans, Int. J. Legal Med. 129 (2015) 963-75.

23. S. Walsh, L. Chaitanya, K. Breslin, C. Muralidharan, A. Bronikowska et al., Global skin colour prediction from DNA. Human Genetics 136 (2017) 847-863.

24. A. R. Wood, T. Esko, J. Yang, S. Vedantam, T.H. Pers et al., Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. 46 (2014) 1173-86.

25. K. Zhong, G. Zhu, X. Jing, A.E.J. Hendriks, S.L.S. Drop et al., Genome-wide compound heterozygote analysis highlights alleles associated with adult height in Europeans. Hum. Genet. 136 (2017) 1407-1417.

26. E. Marouli, M. Graff, C. Medina-Gomez, K.S. Lo, A.R. Wood et al. Rare and low-frequency coding variants alter human adult height. Nature. 9 (2017) 186-190.

27. N. Pirastu, P.K. Joshi, P.S. de Vries, M.C. Cornelis, P.M. McKeigue et al., GWAS for male-pattern baldness identifies 71 susceptibility loci explaining 38% of the risk. Nat. Commun. 8 (2017) 1584.

28. S.P. Hagenaars, W.D. Hill, S.E. Harris, S.J. Ritchie, G. Davies et al., Genetic prediction of male pattern baldness. PLoS Genet. 13 (2017) e1006594.

29. K. Adhikari, T. Fontanil, S. Cal, J. Mendoza-Revilla, M. Fuentes-Guajardo et al., A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. Nat. Commun. 7 (2016) 10815.

30. F. Liu, Y. Chen, G. Zhu, P.G. Hysi, S. Wu et al., Meta-analysis of genome-wide association studies identifies 8 novel loci involved in shape variation of human head hair. Hum. Mol. Genet. (2017) doi: 10.1093/hmg/ddx416.

31. J.R. Shaffer, E. Orlova, M.K. Lee, E.J. Leslie, Z.D. Raffensperger et al., Genome-Wide Association Study Reveals Multiple Loci Influencing Normal Human Facial Morphology. PLoS Genet. 12 (2016) e1006149.

32. J.B. Cole, M. Manyama, E. Kimwaga, J. Mathayo, J.R. Larson et al., Genomewide Association Study of African Children Identifies Association of SCHIP1 and PDE8A with Facial Size and Shape. PLoS Genet. 25 (2016) e1006174.

33. K. Adhikari, M. Fuentes-Guajardo, M. Quinto-Sánchez, J. Mendoza-Revilla, J. Camilo Chacón-Duque et al., A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation. Nat Commun. 7 (2016) 11616.

34. D.J.M. Crouch, B. Winney, W.P. Koppen, W.J. Christmas, K. Hutnik et al., Genetics of the human face: Identification of large-effect single gene variants. Proc Natl Acad Sci U S A. (2018) doi: 10.1073/pnas.1708207114.

35. K. Adhikari, G. Reales, A.J. Smith, E. Konka, J. Palmen et al., A genome-wide association study identifies multiple loci for variation in human ear morphology. Nat. Commun. 6 (2015) 7500.

36. J.R. Shaffer, J. Li, M.K. Lee, J. Roosenboom, E. Orlova et al., Multiethnic GWAS Reveals Polygenic Architecture of Earlobe Attachment. Am. J. Hum. Genet. 101 (2017) 913-924

37. E. Pośpiech, J. Karłowska-Pik, M. Marcińska, S. Abidi, J.D. Andersen et al., Evaluation of the predictive capacity of DNA variants associated with straight hair in Europeans. Forensic Sci. Int. Genet. 19 (2015) 280-288.

38. M. Marcińska, E. Pośpiech, S. Abidi, J. Dyrberg Andersen, M. van den Berge et al., Evaluation of DNA variants associated with androgenetic alopecia and their potential to predict male pattern baldness. PLoS One 10 (2015) e0127852.

39. F. Liu, M.A. Hamer, S. Heilmann, C. Herold, S. Moebus et al., Prediction of male-pattern baldness from genotypes. Eur. J. Hum. Genet. 24 (2016) 895-902.

40. N. Eriksson, J.M. Macpherson, J.Y. Tung, L.S. Hon, B. Naughton et al., Web-based, participant-driven studies yield novel genetic associations for common traits, PLoS Genet. 6 (2010) e1000993.

41. D.H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19 (2009) 1655-64.

42. S.E. Medland, D.R. Nyholt, J.N. Painter, B.P. McEvoy, A.F. McRae et al., Common variants in the trichohyalin gene are associated with straight hair in Europeans, Am J Hum Genet. 85 (2009) 750-5.

43. Tan, Y. Yang, K. Tang, P.C. Sabeti, L. Jin et al., The adaptive variant EDARV370A is associated with straight hair in East Asians, Hum. Genet. 132 (2013) 1187–1191.

44. G. Lubke G, C. Laurin C, R. Walters, N. Eriksson, P. Hysi et al., Gradient Boosting as a SNP Filter: an Evaluation Using Simulated and Hair Morphology Data, J. Data Mining Genomics Proteomics 4 (2013) doi: 10.4172/2153-0602.1000143.

45. M.B. Demay, The hair cycle and Vitamin D receptor. Arch. Biochem. Biophys. 523 (2012) 19-21.

46. H.G. Pálmer, F. Anjos-Afonso, G. Carmeliet, H. Takeda, F.M. Watt, The vitamin D receptor is a Wnt effector that controls hair follicle differentiation and specifies tumor type in adultepidermis. PLoS One 3 (2008) e1483.

47. P. Sriwiriyanont, A. Hachiya, W.L. Pickens, S. Moriwaki, T. Kitahara et al., Effects of IGF-binding protein 5 in dysregulating the shape of human hair. J. Invest. Dermatol. 131 (2011) 320-8.

48. P.M. Steinert, D.A. Parry, L.N. Marekov, Trichohyalin mechanically strengthens the hair follicle: multiple cross-bridging roles in the inner root shealth. J. Biol. Chem. 278 (2003) 41409-19.

49. S. Heilmann, A.K. Kiefer, N. Fricker, D. Drichel, A.M. Hillmer et al., Androgenetic alopecia: identification of four genetic risk loci and evidence for the contribution of WNT signaling to its etiology, J. Invest. Dermatol. 133 (2013) 1489–1496.

50. Y. Shimomura, M. Wajid, Y. Ishii, L. Shapiro, L. Petukhova et al., Disruption of P2RY5, an orphan G protein-coupled receptor, underlies autosomal recessive woolly hair. Nat. Genet. 40 (2008) 335-9.

51. Y. Shimomura, M. Wajid, L. Petukhova, L. Shapiro, A.M. Christiano. Mutations in the lipase H gene underlie autosomal recessive woolly hair/hypotrichosis. J. Invest. Dermatol. 129 (2009) 622-8.

52. Y. Shimomura, M. Wajid, L. Petukhova, M. Kurban, A.M. Christiano, Autosomal-dominant woolly hair resulting from disruption of keratin 74 (KRT74), a potential determinant of human hair texture. Am J Hum Genet 86 (2010) 632-8.

53. A. Fujimoto, M. Farooq, H. Fujikawa, A. Inoue, M. Ohyama et al. A missense mutation within the helix initiation motif of the keratin K71 gene underlies autosomal dominant woolly hair/hypotrichosis. J. Invest. Dermatol. 132 (2012) 2342-9.

54. 1000 Genomes Project Consortium, A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison et al., A global reference for human genetic variation. Nature. 526 (2015) 68-74.

55. H.E. Suchiman, R.C. Slieker, D. Kremer, P.E. Slagboom, B.T. Heijmans et al., Design, measurement and processing of region-specific DNA methylation assays: the mass spectrometry-based method EpiTYPER. Front Genet. 6 (2015) 287.

56. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 26 (2010) 589-95.

57. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan et al., The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25 (2009): 2078-9.

58. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis et al., The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20 (2010) 1297-303.

59. S. Wu, J. Tan, Y. Yang, Q. Peng, M. Zhang et al., Genome-wide scans reveal variants at EDAR predominantly affecting hair straightness in Han Chinese and Uyghur populations. Hum Genet. 135 (2016) 1279-1286.

60. Q. Feng, Y. Lu, X. Ni, K. Yuan, Y. Yang et al., Genetic History of Xinjiang's Uyghurs Suggests Bronze Age Multiple-Way Contacts in Eurasia. Mol. Biol. Evol. 34 (2017) 2572-2582.

61. A. Fujimoto, R. Kimura, J. Ohashi, K. Omi, R. Yuliwulandari, L. Batubara, M.S. et al., A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness, Hum. Mol. Genet. 17 (2008) 835–843.

62. M. Huber, G. Siegenthaler, N. Mirancea, I. Marenholz, D. Nizetic, D. Breitkreutz, D. Mischke, D. Hohl. Isolation and characterization of human repetin, a member of the fused gene family of the epidermal differentiation complex. J. Invest. Dermatol. 124 (2005) 998-1007.

63. E. Pośpiech, S.D. Lee, M. Kukla-Bartoszek, J. Karłowska-Pik, A. Woźniak, M. Boroń, M. Zubańska, A. Bronikowska, S.R. Hong, J.H. Lee, A. Wojas-Pelc, H.Y. Lee, M. Spólnicka, W. Branicki. Variation in the RPTN gene may facilitate straight hair formation in Europeans and East Asians. J Dermatol Sci. 2018 pii: S0923-1811(18)30246-9. doi: 10.1016/j.jdermsci.2018.06.003.

64. B.H. Willis, R.D. Riley. Measuring the statistical validity of summary meta-analysis and meta-regression results for use in clinical practice. Stat Med. 36 (2017) 3283-3301.

65. M. Stone. Cross validatory choice and assessment of statistical predictions. J. Roy. Statist. Soc. Ser. B, 36 (1974) 111-147.

66. R. Kimura R, T. Yamaguchi, M. Takeda, O. Kondo, T. Toma et al., A common variation in EDAR is a genetic determinant ofshovel-shaped incisors. Am. J Hum. Genet. 85 (2009) 528–535.

67. Y.G. Kamberov, S. Wang, J. Tan, P. Gerbault, A. Wark et al., Modeling recent human evolution in mice by expressionof a selected EDAR variant. Cell 152 (2013) 691–702.

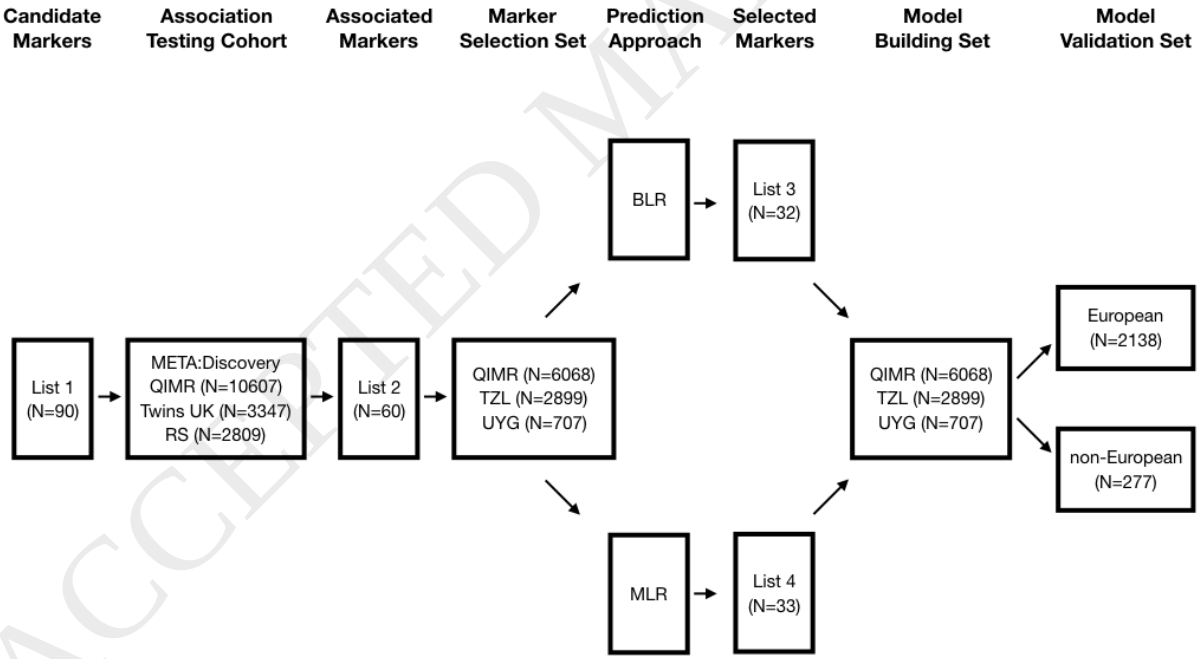
68. R. Zbieć-Piekarska, M. Spólnicka, T. Kupiec, A. Parys-Proszek, Ż. Makowska, A. Pałeczka, K. Kucharczyk, R. Płoski, W. Branicki. Development of a forensically

useful age prediction method based on DNA methylation analysis. Forensic Sci Int Genet. 17 (2015) 173-179.

Figure legends

Figure 1. Overview of the applied study design assuring independent datasets being used for model building and model validation, to avoid overestimation of model outcomes.

Figure 2. Individual contributions of the 32 SNPs to the prediction accuracy expressed by AUC in the BLR model for straight vs. non-straight hair prediction obtained from the model validation set of 2,415 samples (ALL), with 2,138 Europeans (EUR) and 277 non-Europeans (non-EUR). Sequence of SNPs from 1 to 32 is according to Table 3.



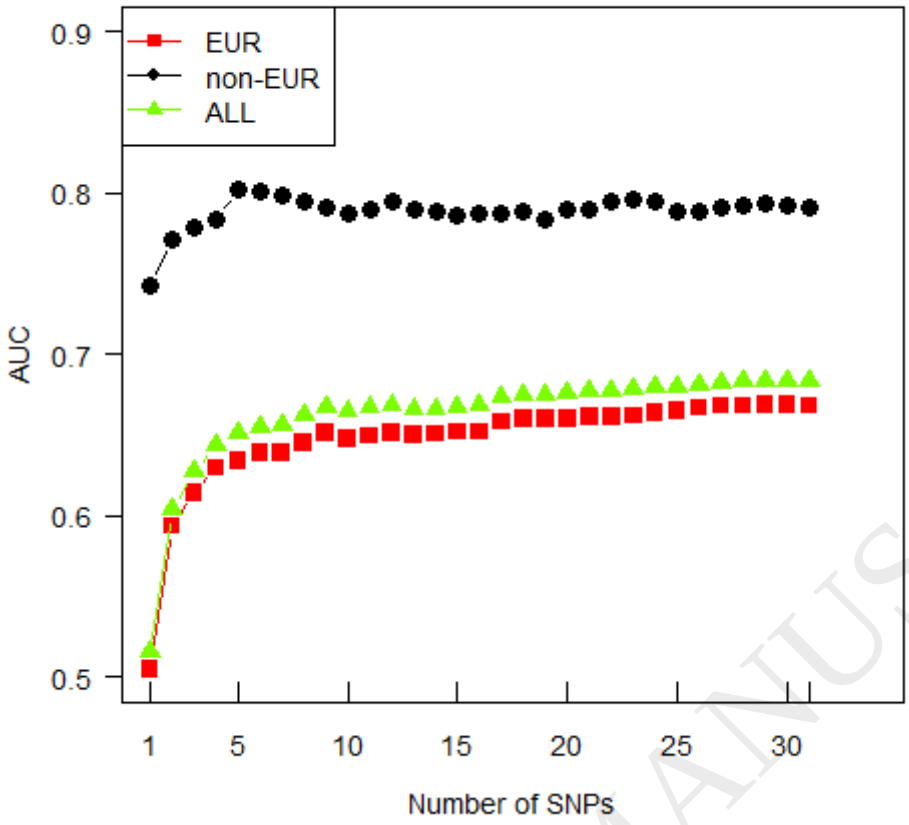


Table 1. Samples and methods employed for genotyping and used as model validation set.

Institution	**Department**	**Genotyping method**	**Phenotyping approach**	**N**
Indiana University Purdue University Indianapolis	Department of Biology	MiSeq	Examination with questionnaire	981
Jagiellonian University*	Malopolska Centre of Biotechnology & Institute of Zoology	Ion Torrent PGM	Clinical examination with interview	628
Erasmus University Medical Centre Rotterdam*	Department of Genetic Identification	Ion Torrent PGM	Questionnaire	128
University of Santiago de Compostela*	Institute of Forensic Sciences, Forensic Genetics Unit	MassARRAY	Examination and interview	117
University of Zurich*	Zurich Institute of Forensic Medicine	Ion Torrent PGM	Interview combined with questionnaire	111
King's College London*	Faculty of Life Sciences and Medicine	MiSeq	Interview	101
University of Copenhagen*	Department of Forensic Medicine, Section of Forensic Genetics,	Ion Torrent PGM	Interview with evaluation of photographs	100
Netherlands Forensic Institute*	Division of Biological Traces	MiSeq	Evaluation of photographs	96
University Hospital Cologne*	Institute of Legal Medicine	Ion Torrent PGM	Photonumeric approach	78
Innsbruck Medical University*	Institute of Legal Medicine	Ion Torrent PGM	Interview	53
University of Münster*	Institute of Legal Medicine	Ion Torrent PGM	Examination and interview	22

*member of EUROFORGEN-NoE Consortium involved in current study

Table 2. Characteristic of samples employed for genotyping and used as model validation set.

Study samples		**N [%]**					**Age**	
		Total	**Straight**	**Wavy**	**Curly**	**Female**	**Mean**	**SD**
European	EUR	2138	870 [40.7]	996 [46.6]	272 [12.7]	1271 [59.4]	36.5	17.8
non-European	Admixed-EUR	61	12 [19.7]	33 [54.1]	16 [26.2]	33 [54.1]	25.0	7.5
	AFR	39	0 [0.0]	2 [5.1]	37 [94.9]	25 [64.1]	23.2	6.2
	AMR	51	15 [29.4]	28 [54.9]	8 [15.7]	37 [72.5]	23.4	9.3
	Middle East	34	6 [17.6]	22 [64.7]	6 [17.6]	18 [52.9]	34.1	10.5
	SAS	50	11 [22.0]	29 [58.0]	10 [20.0]	30 [60.0]	20.6	2.7
	EAS	42	32 [76.2]	9 [21.4]	1 [2.4]	31 [73.8]	22.9	7.1
	Total	277	76 [27.4]	123 [44.4]	78 [28.2]	174 [62.8]	24.5	8.4
ALL		2415	946 [39.2]	1119 [46.3]	350 [14.5]	1445 [59.8]	35.2	17.4

Admixed-EUR: European/non-European admixed, AFR: Africans, AMR: Admixed Americans, SAS: South Asians, EAS: East Asians

SD, Standard Deviation

Table 3. SNPs with their prediction rank from the BLR model to predict straight vs. non-straight hair and the MLR model to predict straight vs. wavy vs. curly hair, respectively in the model building set (N=9,674).

No	SNP ID	Gene	Chromosomal position GRCh37	GWAS-SNP list[a]	Literature Source	Association in META:Discovery[c]	AIC-based prediction rank[d]	
						P value	BLR	MLR
1	rs3827760[e]	*EDAR*[e]	2:109513601		[29,59,61]	-	1	1
2	rs11803731	*TCHH*	1:152083325		[37,40,42]	$2.30x10^{-82}$	2	2
3	rs1268789	*FRAS1*	4:79280693	Yes[b]	[30,37,42]	$4.85x10^{-15}$	3	4
4	rs80293268	*ERRFI1/SLC45A1*	1:8207579	Yes[b]	[30]	$3.66x10^{-9}$	4	5
5	rs310642	*PTK6*	20:62161998	Yes[b]	[30]	$3.74x10^{-10}$	5	6
6	rs1556547	*OFCC1*	6:10270377	Yes[b]	[30,40]	$6.82x10^{-7}$	6	8
7	rs143290289	*KRTAP2-3*	17:39216977	Yes[b]	[30]	$3.71x10^{-8}$	7	9
8	rs11170678	*HOXC13*	12:54154174	Yes[b]	[30]	$1.62x10^{-11}$	8	10
9	rs74333950	*WNT10A*	2:219746292	Yes[b]	[30,42]	$3.98x10^{-15}$	9	7
10	rs10783518	*KRT71*	12:52938497		[52,53]	$2.67x10^{-4}$	10	13
11	rs11203346	*PADI3*	1:17600822	Yes[b]	[30]	$4.58x10^{-8}$	11	12
12	rs1999874	*LINC00708/GATA3*	10:8353101	Yes[b]	[29,30]	$3.72x10^{-9}$	12	11
13	rs6658216	*PEX14*	1:10561604	Yes[b]	[30]	$3.02x10^{-9}$	13	14
14	rs551936	*LIPH*	3:185263467		[51]	$3.81x10^{-3}$	14	17
15	rs12997742	*TGFA*	2:70786598	Yes[b]	[30]	$9.28x10^{-9}$	15	15
16	rs585583	*KRT71*	12:52929370		[52,53,63]	$2.65x10^{-3}$	16	18
17	rs74868796	*HRNR*	1:152191051	Yes	[30]	$3.58x10^{-25}$	17	20
18	rs4845779	1q21.3	1:152479176	Yes	[30]	$4.56x10^{-19}$	18	19
19	rs10788826	1q21.3	1:152161735	Yes	[30]	$2.14x10^{-8}$	19	16
20	rs2219783	*LGR4*	11:27411298	Yes[b]	[30]	$3.84x10^{-8}$	20	21
21	rs9989836	2p14	2:70342727	Yes	[30]	$4.27x10^{-8}$	21	27
22	rs4672907	*LOC391485*	2:219821169		[40,42]	$5.38x10^{-10}$	22	22
23	rs140371183	1q21.3	1:152098428	Yes	[30]	$1.60x10^{-10}$	23	29
24	rs2489250	10p14	10:8274867	Yes	[30]	$6.99x10^{-9}$	24	23
25	rs17646946	*TCHHL1*	1:152062767	Yes[b]	[30,40,42]	$1.78x1^{-84}$	25	31
26	rs3001978	*RPTN*	1:152126467		[40,48,63]	$1.00x10^{-8}$	26	-
27	rs3007671	*NBPF18P/S100A11*	1:151999347		[40,42]	$4.94x10^{-38}$	27	-
28	rs436034	*FRAS1*	4:79256036		[42]	$3.74x10^{-10}$	28	-
29	rs77157375	LINC01494	2:219779911	Yes	[30]	$4.11x10^{-15}$	29	28
30	rs12123907	1q21.3	1:152467751	Yes	[30]	$1.24x10^{-8}$	30	-
31	rs151069963	1q21.3	1:152004241	Yes	[30]	$1.09x10^{-9}$	31	-
32	rs499697	*LCE3E*	1:152493154	Yes[b]	[30,40,42]	$2.57x10^{-17}$	32	33
33	rs11150606	*PRSS53*	16:31099011		[29]	$3.67x10^{-3}$	-	3
34	rs72696935	*TCHH*	1:152085951	Yes	[30]	$2.25x10^{-11}$	-	24
35	rs61816764	*FLG-AS1*	1:152308971	Yes	[30]	$1.27x10^{-10}$	-	25
36	rs2784081	*TRAF2*	9:139791891		[44]	$4.67x10^{-2}$	-	26
37	rs114410520	LOC105371441	1:151972609	Yes	[30]	$4.17x10^{-10}$	-	30
38	rs11582331	*RPTN*	1:152134136		[40,48,63]	$3.10x10^{-5}$	-	32

[a]SNPs selected from our recently published large GWAS and meta-analysis [30]
[b]SNPs included in a 14-SNP prediction model reported in [30]; for 3 SNPs (rs506863 in *FRAS1*, rs11078976 in *KRTAP*, and rs2847344 in *PEX14*) substitute LD SNPs (rs1268789 in *FRAS1*, rs143290289 in *KRTAP2-3* and rs6658216 in *PEX14*) were used due to technical difficulties
[c]from GWAS-meta-analysis of 16,763 Europeans described in [30]
[d]using 9,674 Europeans and East Asians (6068 QIMR, 2899 TZL, 707 UYG) from the model building set
[e]rs3827760 was not covered by the META:Discovery because of European subjects used in this GWAS
BLR, binomial logistic regression; MLR, multinomial logistic regression

Table 4. Accuracy estimates of the BLR model to predict head hair shape as straight vs. non-straight from 32 SNPs with and without considering sex and age, obtained from the model validation set.

BLR model	Sample set	AUC	Sensitivity	Specificity	PPV	NPV
32 SNP predictors without sex and age	ALL (N=2,415)	0.679	0.840	0.364	0.460	0.780
	EUR (N=2,138)	0.664	0.841	0.342	0.467	0.759
	non-EUR (N=277)	0.789	0.829	0.498	0.384	0.885
32 SNP predictors with sex and age	ALL (N=2,415)	0.695	0.844	0.364	0.461	0.783
	EUR (N=2,138)	0.680	0.846	0.338	0.467	0.762
	non-EUR (N=277)	0.800	0.816	0.527	0.395	0.883

Table 5. Accuracy estimates of the MLR model to predict head hair shape in 3 categories straight vs. wavy vs. curly from 33 SNPs as obtained from the model training set.

MLR model	Sample set	Hair shape	AUC	Sensitivity	Specificity	PPV	NPV
33 SNP predictors without sex and age*	ALL (N=2,415)	Straight	0.680	0.860	0.340	0.456	0.791
		Wavy	0.597	0.010	0.995	0.611	0.538
		Curly	0.621	0.006	0.998	0.286	0.855
	EUR (N=2,138)	Straight	0.666	0.862	0.324	0.467	0.774
		Wavy	0.596	0.009	0.996	0.643	0.535
		Curly	0.600	0.004	0.997	0.167	0.873
	non-EUR (N=277)	Straight	0.801	0.842	0.438	0.362	0.880
		Wavy	0.609	0.016	0.987	0.500	0.557
		Curly	0.736	0.013	1.000	1.000	0.721

*for the model with sex and age, see Supplementary Table 6